

# A Factorized Model for Transitive Verbs in Compositional Distributional Semantics

**Lilach Edelstein**

ELSC

Hebrew University of Jerusalem

[lilach.edelstein@mail.huji.ac.il](mailto:lilach.edelstein@mail.huji.ac.il)

**Roi Reichart**

Faculty of IE&M

Technion, IIT

[roiri@ie.technion.ac.il](mailto:roiri@ie.technion.ac.il)

## Abstract

We present a factorized compositional distributional semantics model for the representation of transitive verb constructions. Our model first produces (subject, verb) and (verb, object) vector representations based on the similarity of the nouns in the construction to each of the nouns in the vocabulary and the tendency of these nouns to take the subject and object roles of the verb. These vectors are then combined into a final (subject, verb, object) representation through simple vector operations. On two established tasks for the transitive verb construction our model outperforms recent previous work.

## 1 Introduction

In recent years, vector space models, deriving word meaning representations from word co-occurrence patterns in text, have become prominent in lexical semantics research (Turney et al., 2010; Clark, 2012). Following this success recent attempts have been devoted to compositional distributional semantics (CDS): combining the distributional word representations, often in a syntax-driven fashion, to produce representations of phrases and sentences.

Several tasks and techniques have been proposed for CDS. Some work aims at representing sentences that vary in length and structure mostly with neural network models (Socher et al., 2012; Marelli et al., 2014; Le and Mikolov, 2014; Pham et al., 2015). Another approach, which we take in this paper, is to focus on specific syntactic constructions. At the expense of

generality, this approach enables an in-depth investigation of a specific linguistic phenomenon. Mitchell and Lapata (2008) proposed various additive and multiplicative operators for the combinations of word vectors, and applied them to intransitive verbs and their subjects. Recently, the categorical framework (Coecke et al., 2011; Baroni et al., 2014) has been proposed, where each word is represented by a tensor whose order is determined by the categorical grammar type of the word. For example, Baroni and Zamparelli (2010) represent nouns by a vector, and adjectives by matrices transforming one noun vector into another.

In this paper we focus on the transitive verb construction, which recently attracts much attention. In the categorical framework it is represented with a third order tensor that takes the noun vectors representing the subject and object and returns a vector in the sentence space (Grefenstette et al., 2013; Polajnar et al., 2014). The main limitation of this approach is the excessive number of involved parameters. For example, a third-order tensor for a given transitive verb, mapping two 100-dimensional noun spaces to a 100-dimensional sentence space, would have  $100^3$  parameters in its full form. Indeed, several recent works have tried to reduce the size of these models (Polajnar et al., 2014; Fried et al., 2015) while others proposed matrix based representations (Polajnar et al., 2014; Milajevs et al., 2014; Paperno et al., 2014).

We propose a *factorized model* for the representation of transitive verb constructions. Given a subj-verb-obj  $(s, v, o)$  construction, our model

builds vector representations for the  $(s, v)$  and the  $(v, o)$  pairs, based on the similarity of  $s$  and  $o$  to each of the nouns in the vocabulary and the tendency of these nouns to take the subject and object roles of  $v$ . The dimensionality of these  $(s, v)$  and  $(v, o)$  vectors (15701 in our case) equals to the number of nouns in the vocabulary that take the subject and object positions of  $v$  frequently enough. The  $(s, v)$  and  $(v, o)$  vectors are then combined to a final  $(s, v, o)$  vector through simple vector operations. Our model outperforms recent previous work on two established tasks for the transitive verb construction (Grefenstette and Sadrzadeh, 2011b; Kartsaklis et al., 2014).

## 2 Model

The goal of our model is to generate vector representations (embeddings) for subject-verb-object  $(s, v, o)$  constructions, where  $s$  is the subject noun,  $v$  is a verb and  $o$  is the object noun. The model consists of two steps: **(1)** Embed the  $(s, v)$  and  $(v, o)$  pairs based on co-occurrence statistics of the members of each pair with all other nouns in the vocabulary; and **(2)** Combine the  $(s, v)$  and  $(v, o)$  representations to create a final  $(s, v, o)$  embedding.

Our model is a *factorized* model, generating an  $(s, v, o)$  representation from its pair components  $((s, v)$  and  $(v, o))$ . As such it is compact: the size of the  $(s, v)$  and  $(v, o)$  vectors is the number of nouns in the vocabulary that appear frequently enough both as subjects and as objects of  $v$ , and the  $(s, v, o)$  vector is a simple derivation of these vectors. In what follows we describe each of the above steps.

### 2.1 Pair Representations

We represent the  $(s, v)$  and  $(v, o)$  pairs through the relations between the verb ( $v$ ) and the noun ( $s$  or  $o$ ) with each other noun in the vocabulary. Particularly, we consider the tendency of each noun to come as a subject (for  $(s, v)$ ) or object (for  $(v, o)$ ) of  $v$ , and the similarity of that noun to  $s$  or  $o$ , respectively. By considering all the nouns in the vocabulary we get a smooth estimate of the tendency of the noun ( $s$  or  $o$ ) to take the subject (for  $s$ ) or object (for  $o$ ) position of  $v$ . In what follows, we describe the representation in details for  $(s, v)$  pairs. A very similar process is employed for  $(v, o)$  pairs.

For an  $(s, v)$  pair we construct a vector representation whose size is the number of nouns that appear frequently enough at both subject and object positions in the training corpus. The  $k$ 'th coordinate in this representation is given by:

$$u_{s,v}^{subj}[k] = NVSubj(n_k, v) \cdot NNSim(n_k, s)$$

where  $n_k$  is the  $k$ 'th noun in the vocabulary,  $NVSubj(x, y)$  reflects the tendency of the noun  $x$  to be the subject of the verb  $y$ , and  $NNSim(x, y)$  reflects the similarity between the nouns  $x$  and  $y$ . We next describe how  $NVSubj(x, y)$  and  $NNSim(x, y)$  are computed.

**NVSubj** For a noun  $x$  and a verb  $y$ , we compute the positive point-wise mutual information (PPMI) of  $x$  appearing as the subject of  $y$ , and the score for the  $(x, y)$  pair is then given by:

$$NVSubj(x, y) = PPMI^{subj}(x, y) = \max(0, \log((P^{subj,verb}(x, y)) / (P^{subj}(x)P^{verb}(y))))$$

where  $P^{subj,verb}(x, y)$  is the probability that a (subject,verb) pair in the corpus is  $(x, y)$ , and  $P^{subj}(x)$  and  $P^{verb}(y)$  are the probabilities that  $x$  appears at the subject position of any verb in the corpus and that  $y$  appears as a verb in the corpus, respectively.

**NNSim** This score reflects the similarity between two nouns:  $NNSim(x, y) = sim(x, y)$ , where  $sim(x, y)$  is any function that returns the similarity between its two word arguments.

We apply the same considerations for  $(v, o)$  pairs: for the  $k$ -th coordinate,  $NVObj$  represents the tendency of  $n_k$  to be an object of  $v$ , while  $NNSim$  represents the similarity between  $n_k$  and  $o$ .

### 2.2 $(s, v, o)$ Construction Representation

Given the  $(s, v)$  and  $(v, o)$  representations described above, our next step is to combine them so that to get an effective representation of the  $(s, v, o)$  triplet.

We consider three combination methods. In the first two a single vector is constructed for  $(s, v, o)$  through: (1) concatenation of the two vectors; and (2) coordination-wise multiplication:

$$u^{(s,v,o)}[k] = u_{s,v}^{subj}[k] \cdot u_{v,o}^{obj}[k] = NVSubj(n_k, v) \cdot NNSim(n_k, s) \cdot NVObj(n_k, v) \cdot NNSim(n_k, o)$$

That is, the  $k$ -th coordinate represents the similarity of  $n_k$  to both  $s$  and  $o$ , and its co-occurrence statistics with  $v$  as both a subject and an object. Under these two combination methods the similarity score of two  $(s, v, o)$  constructions is defined to be the cosine similarity between their vectors.

As an alternative, the third combination method keeps the  $(s, v)$  and  $(v, o)$  vectors as a representation of  $(s, v, o)$ . The similarity between two constructions  $(s, v, o)^1$  and  $(s, v, o)^2$  is computed through the similarities between their components:

$$\text{sim}((s, v, o)^1, (s, v, o)^2) = \text{cosine}(u_{(s,v)^1}, u_{(s,v)^2}) \cdot \text{cosine}(u_{(v,o)^1}, u_{(v,o)^2})$$

### 3 Experiments

**Data Preprocessing and Training** We trained our models on the cleaned and tokenized Polyglot Wikipedia corpus (Al-Rfou et al., 2013),<sup>1</sup> consisting of approximately 75M sentences and 1.5G word tokens. The corpora were POS-tagged with universal POS (UPOS) tags (Petrov et al., 2012) using the TurboTagger (Martins et al., 2013),<sup>2</sup> trained with default settings (SVM MIRA with 20 iterations) without any further parameter fine-tuning, on the TRAIN+DEV portion of the UD treebank. Following, the corpus was parsed with Universal Dependencies<sup>3</sup> using the Mate parser v3.61 (Bohnet, 2010),<sup>4</sup> trained on the same UD treebank portion as the tagger and with default settings.

After parsing the corpus and before further statistics were collected, the corpus was lemmatized to facilitate robust estimation. Therefore we also considered the lemmas of the words in our evaluation sets when computing an  $(s, v, o)$  representation. We extracted all  $(n, v)$  and  $(v, o)$  pairs based on dependency labels: a noun or a pronoun modifying a verb were considered its subject if their dependency arc is labeled "subj" or "nsubjpass", and its object if their dependency arc is labeled "dobj", "iobj", "nmod" or "xcomp". In order to reduce sparsity, our vocabulary contains only verbs that appear at least 50 times in the corpus and nouns that appear at least 50 times

both at subject and at object positions, a total of 6934 transitive verbs and 15701 nouns.

To compute the similarity between two nouns with *NNSim*, we trained the word2vec skip-gram model with negative sampling (Mikolov et al., 2013) on our (unparsed) training corpus; context-window size was set to 5 and vector dimensionality to 200.

**Evaluation** We evaluate the performance of our models on two well established tasks for transitive verb constructions. Both tasks require ranking of transitive sentence pairs for semantic similarity. The gold standard ranking is derived from similarity scores, on a 1-7 scale, provided by human evaluators. The model ranking is evaluated against the gold standard ranking using Spearman's  $\rho$ .

The first task (GS11, (Grefenstette and Sadrzadeh, 2011b)) involves verb disambiguation: each of the 200 pairs in the dataset consists of two sentences that differ in their transitive verb, but share the same subject and object. For example, the members of the pair "(man, draw, sword), (man, attract, sword)" are less similar than those of "(report, draw, attention), (report, attract, attention)".

The second task uses the transitive sentence similarity dataset (KS14, (Kartsaklis et al., 2014)). This dataset consists of 108 subject-verb-object pairs, derived from 72 subject-verb-object triplets arranged into pairs. Unlike GS11, here each pair is composed of two triplets that differ in all three words. For example, the pair "(programme, offer, support), (service, provide, help)" is expected to get a higher similarity score compared to "(school, encourage, child), (employee, leave, company)".

In both tasks, we consider two different aggregation methods over the annotator scores of a pair: (a) the human scores of each annotator are paired with the model scores without averaging, and a  $\rho$  score is computed between the two vectors; and (b) the human scores are first averaged, a human ranking is derived from the averaged scores and compared to the model ranking. While the second method may seem more robust, it was not used in most previous works (see discussion in Mitchell and Lapata (2008)).

**Models and Baselines** We compare the results of our models, distinguished by the three combi-

<sup>1</sup><https://sites.google.com/site/rmyeid/projects/polyglot>

<sup>2</sup><http://www.cs.cmu.edu/ark/TurboParser/>

<sup>3</sup>We use UD so that in future work we can apply our model to other languages without major language-specific adaptations.

<sup>4</sup><https://code.google.com/archive/p/mate-tools/>

	Averaged		Non-Averaged	
	GS11	KS14	GS11	KS14
w2v-sum	0.280	<b>0.760</b>	0.210	<b>0.60</b>
w2v-all	0.473	0.716	0.415	0.566
2014-best-non-simple	0.456	0.655	–	–
2014-best-simple	0.348	0.732	–	–
2015-best-tensor	0.470	0.680	0.360	0.520
2015-best-simple	–	0.710	0.140	0.560
w2v-PPMI-concat	0.601	0.743	<b>0.457</b>	0.562
w2v-PPMI-coor-mult	0.598	0.716	0.454	0.567
w2v-PPMI-mult-score	<b>0.605</b>	0.653	0.454	0.564

**Table 1:** Spearman  $\rho$  scores between model and human rankings for the GS11 and the KS14 tasks with averaged and non-averaged human scores. Top two models: baselines. Next four models: previous work. Bottom three models: this paper.

nation methods: concatenation (w2v-PPMI-concat), coordination-wise multiplication (w2v-PPMI-coor-mult) and multiplication of the  $(s, v)$  and  $(v, o)$  scores (w2v-PPMI-mult-score). We consider several baselines. In a first, simple baseline (w2v-sum), each  $(s, v, o)$  construction is represented as the sum of the word2vec vectors of its words. This baseline, which corresponds to the unsupervised additive method of Mitchell and Lapata (2008), captures the strength of the word2vec word level representations, ignoring word order and syntactic structure considerations.<sup>5</sup>

A second baseline (w2v-all) is similar to our method but the syntactic information is replaced with word similarity based on word2vec scores:

$$u_{s,v}^{subj}[k] = NVSim(n_k, v) \cdot NNSim(n_k, s)$$

$$u_{v,o}^{obj}[k] = NVSim(n_k, v) \cdot NNSim(n_k, o)$$

Where  $NVSim(x, y)$  is the cosine similarity between the word2vec vectors of  $x$  and  $y$ . This method quantifies the importance syntax to our model.

Finally, we compare to state-of-the-art previous work: (a) the most recent study on our tasks ((Fried et al., 2015), their table 1, *2015-best-tensor*: their best model; *2015-best-simple*: best result with additive or multiplicative combination as in Mitchell and Lapata (2008)); and (b) Milajevs et al. (2014) which performed exhaustive comparison of models and vector representations for our tasks (see their table 2, *2014-best-simple*: best result with additive or

multiplicative combination; *2014-best-non-simple*: best result with tensor and matrix combinations based on (Grefenstette and Sadrzadeh, 2011a; Grefenstette and Sadrzadeh, 2011b; Kartsaklis et al., 2012; Kartsaklis et al., 2014)).<sup>6</sup>

## 4 Results

Results are presented in Table 1. Our models are superior on the GS11 task: the gaps between our best model and the best baseline are 13.2 and 4.2  $\rho$  points for the averaged and the non-averaged conditions, respectively. When comparing to the best previous work the gaps are 13.5 and 9.7  $\rho$  points, respectively.

For the KS14 task it is the simple w2v-sum baseline that performs best ( $\rho = 0.76$  for averaged scores,  $\rho = 0.6$  for non-averaged scores), but in both conditions it is one of our models that is second best (w2v-PPMI-concat with  $\rho = 0.743$  for averaged scores, w2v-PPMI-coor-mult with  $\rho = 0.567$  for non-averaged scores). Yet, in this task our gap from the baselines are smaller compared to GS11.

The superiority of a simple additive model for KS14 is in line with previously reported results. While in KS14 the compared  $(s, v, o)$  constructions do not overlap in their lexical content, in GS11 paired constructions differ only in the verb, hence requiring finer grained distinctions. The relative difficulty of GS11 is also reflected by the lower scores all participating models achieve on this task.

Importantly, while the w2v-sum excels on the KS14 task, its performance substantially degrade on the GS11 task where it achieves  $\rho$  values of only 0.28 and 0.21 for the averaged and non-averaged cases respectively. Our models hence provide a sweet spot of good performance on both tasks.

Finally, the three variants of our model perform very similarly in three out of four test conditions. It is only for KS14 with averaged human scores that w2v-PPMI-mult-score lags 5.3 and 9  $\rho$  points behind w2v-PPMI-coor-mult and w2v-PPMI-concat, respectively. Hence, our model is flexible with respect to combination method selection.

<sup>6</sup>Note that \*-tensor, \*-simple and \*-non-simple do not necessarily refer to the same model at all conditions: we pick the best result for each task and human scoring combination among the models in each of these model groups.

<sup>5</sup>We do not report results with coordination-wise multiplication of w2v vectors, as they lag behind other reported models.



## 5 Conclusions

We presented a factorized model for  $(s, v, o)$  embeddings which provides a simple and compact alternative to existing methods, and showed that it excels on two recent CDS tasks. In future work we intend to extend our model so that it accounts for more complex syntactic constructions, such as, e.g., ditransitive verb constructions and constructions that include adjectives and adverbs.

## References

- [Al-Rfou et al.2013] Rami Al-Rfou, Bryan Perozzi, and Steven Skiena. 2013. Polyglot: Distributed word representations for multilingual NLP. In *Proc. of CoNLL*.
- [Baroni and Zamparelli2010] Marco Baroni and Roberto Zamparelli. 2010. Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space. In *Proc. of EMNLP*.
- [Baroni et al.2014] Marco Baroni, Raffaella Bernardi, and Roberto Zamparelli. 2014. Frege in space: A program of compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.
- [Bohnet2010] Bernd Bohnet. 2010. Top accuracy and fast dependency parsing is not a contradiction. In *Proc. of COLING*.
- [Clark2012] Stephen Clark. 2012. Vector space models of lexical meaning. *Handbook of Contemporary Semantics*, Wiley-Blackwell, à paraître.
- [Coecke et al.2011] Bob Coecke, Mehrnoosh Sadrzadeh, and Stephen Clark. 2011. Mathematical foundations for a compositional distributional model of meaning. *Linguistic Analysis*, 36(1-4):345–384.
- [Fried et al.2015] Daniel Fried, Tamara Polajnar, and Stephen Clark. 2015. Low-rank tensors for verbs in compositional distributional semantics. In *Proc. of ACL (short papers)*.
- [Grefenstette and Sadrzadeh2011a] Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011a. Experimental support for a categorical compositional distributional model of meaning. In *Proc. of EMNLP*.
- [Grefenstette and Sadrzadeh2011b] Edward Grefenstette and Mehrnoosh Sadrzadeh. 2011b. Experimenting with transitive verbs in a discocat. In *Proceedings of the GEMS 2011 Workshop on GEometrical Models of Natural Language Semantics*.
- [Grefenstette et al.2013] Edward Grefenstette, Georgiana Dinu, Yao-Zhong Zhang, Mehrnoosh Sadrzadeh, and Marco Baroni. 2013. Multi-step regression learning for compositional distributional semantics. In *Proc. of IWCS*.
- [Kartsaklis et al.2012] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Stephen Pulman. 2012. A unified sentence space for categorical distributional-compositional semantics: Theory and experiments. In *Proc. of COLING*.
- [Kartsaklis et al.2014] Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, Chris Heunen, Manuel L Reyes, Ravi Kunjwal, and Tobias Fritz. 2014. A study of entanglement in a categorical framework of natural language. In *Proceedings of the 11th Workshop on Quantum Physics and Logic (QPL 2014)*.
- [Le and Mikolov2014] Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *Proc. of ICML*.
- [Marelli et al.2014] Marco Marelli, Stefano Menini, Marco Baroni, Luisa Bentivogli, Raffaella Bernardi, and Roberto Zamparelli. 2014. A sick cure for the evaluation of compositional distributional semantic models. In *Proc. of LREC*.
- [Martins et al.2013] André F. T. Martins, Miguel B. Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proc. of ACL (short papers)*.
- [Mikolov et al.2013] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.
- [Milajevs et al.2014] Dmitrijs Milajevs, Dimitri Kartsaklis, Mehrnoosh Sadrzadeh, and Matthew Purver. 2014. Evaluating neural word representations in tensor-based compositional settings. In *Proc. of EMNLP*.
- [Mitchell and Lapata2008] Jeff Mitchell and Mirella Lapata. 2008. Vector-based models of semantic composition. In *Proc. of ACL*.
- [Paperno et al.2014] Denis Paperno, Nghia The Pham, and Marco Baroni. 2014. A practical and linguistically-motivated approach to compositional distributional semantics. In *Proc. of ACL*.
- [Petrov et al.2012] Slav Petrov, Dipanjan Das, and Ryan T. McDonald. 2012. A universal part-of-speech tagset. In *In Proc. of LREC*.
- [Pham et al.2015] Nghia The Pham, Germán Kruszewski, Angeliki Lazaridou, and Marco Baroni. 2015. Jointly optimizing word representations for lexical and sentential tasks with the c-phrase model. In *Proc. of ACL*.
- [Polajnar et al.2014] Tamara Polajnar, Luana Fagarasan, and Stephen Clark. 2014. Reducing dimensions of tensors in type-driven distributional semantics. In *Proc. of EMNLP*.
- [Socher et al.2012] Richard Socher, Brody Huval, Christopher D Manning, and Andrew Y Ng. 2012. Semantic compositionality through recursive matrix-vector spaces. In *Proc. of EMNLP-CoNLL*.

[Turney et al.2010] Peter D Turney, Patrick Pantel, et al.  
2010. From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37(1):141–188.